Report for 2002MI5B: Grid Computing for Real Time Distributed Collaborative Geoprocessing with Applications in Water Quality Management

- Articles in Refereed Scientific Journals:
    - Shi, Y, Shortridge, A. M., and Bartholic, J. F. 2002. Grid Computing for Real Time Distributed Collaborative Geoprocessing, in Advances in Spatial Data Handling: Proceedings of the 10th International Symposium on Spatial Data Handling, Springer Verlag; ISBN: 3540438025; July); Bk&Cd-Rom edition (September 2002).

Report Follows:

Title: Grid Computing for Real Time Distributed Collaborative Geoprocessing--with Applications in Water Quality Management
Project Number: 2002MI4B
Start Date: 03/01/2001
End Date: 02/28/2002
Funding Source: US Geological Survey
Congressional District: Eighth
Lead Institute: Michigan Institute of Water Research
Descriptors: Water Quality; Watershed Management; Web based GIS; Monitoring
Primary PI: Yi Shi, Institute of Water Research, Michigan State University

Problem and Research Objective(s)
Grid computing has emerged as an important new field in the distributed computing arena (Foster et al., 2001; Foster & Kesselman, 1997). It focuses on large-scale resource sharing, innovative applications and, in some cases, high-performance orientation within a so-called Virtual Organization (VO). It has great capability to link multiple agencies with a network of shared data, software, and processors. The degree of integration that a VO Grid can achieve employing computing technologies is unprecedented. The hypothetical benefits of grid computing for spatial processing, in particular hydrologic applications, appear great. However we are unaware of any research exploring this new technology's potential. Distributed computing itself is not a new topic in geographic information science (Coleman et al., 1999; UCGIS, 1998). Parallelization of spatial operations has also been an area of active research (Magillo & Puppo, 1998; Hoel & Samet, 1996). We maintain that integrating the two topics via grid computing shows great promise for enhancing a variety of geospatial applications, particularly those with intensive computing requirements and a multi-organizational structure. Hydrologic modeling and water quality management stand to benefit tremendously from this integration.

We are conducting experiments on a water quality management VO testbed. Water quality management is a holistic activity that involves coordination among different organizations and collective decision-making based upon information from different sources. The ultimate goal is to model agricultural non-point source water contamination as part of an effort to improve water quality and support land use planning and agricultural production on a sustained basis. There is also an urgent need to educate the general public on water quality problems. Due to its organizational characteristics and modeling goals, water quality management is an excellent theme upon which to develop a VO/Grid computing testbed. This is a problem-oriented framework that contains data and methods for facilitating decision-making in a particular geographic region of a watershed. Our experiments are directed to the challenging task of distributing data and processing for a real time pollutant loading model across the network.

Methodology
Modern distributed and component GIS technologies and advanced distributed computing techniques are heavily used in this research. We draw upon our current experience with grid computing for basic hydrologic functionality to accomplish the more advanced objectives of this proposal.

First, we designed a general Internet-based integration framework for watershed management based on VO/Grid computing technologies using existing computing facilities. Then we created an example water quality management system that utilizes this framework. Hydrologic modeling algorithms were paralleled and datasets were partitioned to enable processing distribution across the Internet. Basic functionality for this system was developed using the MPI libraries and C++ software written by the investigators.

Principle Findings and Significance

In this research Grid, computing was described as an important recent development in the general area of distributed computing. This is due to its ability to link multiple agencies with a network of shared data, software, and processors. The innovation is the degree of integration that a VO employing such technologies can achieve.

We identified three major driving factors for the implementation of Grid-based geoprocessing. As network speeds continue to increase relative to single processing power, parallel implementations become relatively efficient. Second spatial partitioning for local operators is simple and very general. Partitioning for neighborhood operations ranges from simple (slope) to complex (flow accumulation). Complex partitioning problems restrict the utility of parallel implementations. Finally, algorithm complexity is a factor favoring parallelization, since users can take advantage of more powerful processors than are available on the local network.

The trend toward large-scale multi-agency geospatial projects involving the collaboration of dozens of scientists, policy-makers, and members of the public makes the virtual organization paradigm attractive. This research identified the utility of real-time distributed model processing for such organizations. We maintain that Grid technologies show great promise for enhancing a variety of geospatial applications, particularly those with intensive computing requirements and a multi-organizational structure.

Cited References

Coleman, D.J. (1999) "Geographical information systems in networked environments", in Longley, P.A., Goodchild, M.F., Maguire, D.J., and Rhind, D.W. (eds.), Geographic Information Systems, New York: Wiley, 317-329.

Foster, I., Kesselman, C., and Tuecke, S. (2001) "The anatomy of the Grid: enabling scalable virtual organizations", Intl. J. Supercomputer Applications (in press).

Foster, I., and Kesselman, C. (1997) "Globus: A Metacomputing Infrastructure Toolkit", Intl. J. Supercomputer Applications 11(2), 115-128.

Hoel, E.G., and Samet, H. (1994) "Performance of data-parallel spatial operations", Proceedings 20th Intl. Conf. On Very Large DataBases, September 1994, Santiago, Chile, 156-167.

Magillo, P., and Puppo, P. (1998) "Algorithms for parallel terrain modeling and visualization", in Healey, R., Dowers, S., Gittings, B., and Mineter, M. (eds.) Parallel processing algorithms for GIS, London: Taylor & Francis, 351-386.

UCGIS (1998) "Distributed Computing", Revised UCGIS White Papers. Online: http://www.ucgis.org/research_white/distrib.html, last checked 10/15/01.